Supplemental Materials

In the paper submitted to ICDM 2016, we proposed to train the model via variational inference. However, we realized that an Expectation Maximization (EM) algorithm [1] that finds a Maximum a Posteriori (MAP) estimator is simpler and more efficient. Therefore, we present the EM algorithm here.

In the EM algorithm, we treat $\mathbf{Z}$ as the latent variables and find a MAP estimator for the parameters $\mathbf{T}$ and $\boldsymbol{\Theta}$. The objective function is given by

$$\begin{aligned}\widehat{\mathbf{T}}, \widehat{\boldsymbol{\Theta}} &= \arg\max_{\mathbf{T},\boldsymbol{\Theta}} \log p(\mathbf{T}, \mathbf{Y}, \boldsymbol{\Phi} | \mathbf{X}, \boldsymbol{\Theta}) \\ &= \arg\max_{\mathbf{T},\boldsymbol{\Theta}} \log \sum_{\mathbf{Z}} p(\mathbf{T}, \mathbf{Y}, \mathbf{Z}, \boldsymbol{\Phi} | \mathbf{X}, \boldsymbol{\Theta}) \\ &= \arg\max_{\mathbf{T},\boldsymbol{\Theta}} \log \sum_{\mathbf{Z}} \Big\{ \prod_{k=1}^{K} \prod_{d=1}^{D} p(\mathbf{t}_{kd} | \alpha_t, \beta_t, \mu_{t_{kd}^-}, \mu_{t_{kd}^+}) \\ &\quad \prod_{n=1}^{N} p(\mathbf{z}_n) \prod_{n=1}^{N} p(\phi_n | \mathbf{z}_n, \mathbf{x}_n, \mathbf{T}) \prod_{n=1}^{N} p(\mathbf{y}_n | \mathbf{z_n}, \boldsymbol{\Theta}) \Big\}\end{aligned} \tag{1}$$

In the EM algorithm, we iteratively apply the Expectation step (E Step) and Maximization Step (M Step) until convergence.

Expectation Step

In the E Step, we first compute the posterior distribution of the latent variable $\mathbf{Z}$, given the parameters $\{\mathbf{T}, \boldsymbol{\Theta}\}$ and the observed variables $\{\boldsymbol{\Phi}, \mathbf{X}, \mathbf{Y}\}$. The log posterior distribution is given by

$$\begin{aligned}&\log p(\mathbf{Z} | \mathbf{T}, \boldsymbol{\Theta}, \boldsymbol{\Phi}, \mathbf{X}, \mathbf{Y}) \\ &= \sum_{n=1}^{N} \Big\{ \log p(\mathbf{z}_n) + \log p(\phi_n | \mathbf{z}_n, \mathbf{x}_n, \mathbf{T}) + \log p(\mathbf{y}_n | \mathbf{z}_n, \boldsymbol{\Theta}) \Big\} + const \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \Big\{ \sum_{d=1}^{D} \log g(x_{nd} - t_{kd}^-) + \log g(t_{kd}^+ - x_{nd}) \\ &\quad - \log \Big(1 - \prod_{d=1}^{D} g(x_{nd} - t_{kd}^-) g(t_{kd}^+ - x_{nd}) \Big) \\ &\quad - \frac{1}{2} \log |\boldsymbol{\Sigma}_k| - \frac{1}{2} (\mathbf{y}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{y}_n - \boldsymbol{\mu}_k) \Big\} + const\end{aligned} \tag{2}$$

where $const$ represents constants that are not functions of $\mathbf{Z}$ and normalize this equation such that it defines a valid probability. By observing this Equation, we conclude that the posterior distribution for each $\mathbf{z}_n$ with $n \in \{1 \dots N\}$ is conditional independent with each other, and is given as

$$\mathbf{z_n} | \mathbf{x}_n, \phi_n, \mathbf{y}_n, \mathbf{T}, \boldsymbol{\Theta} \sim \text{Categorical}(\boldsymbol{\pi}_n) \tag{3}$$

where $\boldsymbol{\pi}_n$ is a K-dimensional vector, each of whose element is defined by

$$\begin{aligned}\pi_{nk} \propto \exp \Big\{ &\sum_{d=1}^{D} \log g(x_{nd} - t_{kd}^-) + \log g(t_{kd}^+ - x_{nd}) \\ &- \log \Big(1 - \prod_{d=1}^{D} g(x_{nd} - t_{kd}^-) g(t_{kd}^+ - x_{nd}) \Big) \\ &- \frac{1}{2} \log |\boldsymbol{\Sigma}_k| - \frac{1}{2} (\mathbf{y}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{y}_n - \boldsymbol{\mu}_k) \Big\}\end{aligned} \tag{4}$$

and $\boldsymbol{\pi}_n$ is normalized such that $\sum_{k=1}^{K} \pi_{nk} = 1$. The expected value of z_{nk} with respect to the posterior distribution is given as $\mathbb{E}[z_{nk}] = \pi_{nk}$.

Given the posterior distribution of $\mathbf{Z}$, we are able to compute the expected value of the log joint distribution, denoted by $Q(\mathbf{T}, \boldsymbol{\Theta})$ such that

$$\begin{aligned}Q(\mathbf{T}, \boldsymbol{\Theta}) =& \mathbb{E}[\log p(\mathbf{T}, \mathbf{Y}, \mathbf{Z}, \boldsymbol{\Phi} | \mathbf{X}, \boldsymbol{\Theta})] \\ =& -\frac{1}{2}\alpha_t \sum_{k=1}^{K} \sum_{d=1}^{D} (t_{kd}^- - \mu_{t_{kd}^-})^2 - \frac{1}{2}\alpha_t \sum_{k=1}^{K} \sum_{d=1}^{D} (t_{kd}^+ - \mu_{t_{kd}^+})^2 - \frac{1}{2}\beta_t \sum_{k=1}^{K} \sum_{d=1}^{D} (t_{kd}^+ - t_{kd}^-)^2 \\ &+ \sum_{n=1}^{N} \sum_{k=1}^{K} \pi_{nk} \sum_{d=1}^{D} \Big(\log g(x_{nd} - t_{kd}^-) + \log g(t_{kd}^+ - x_{nd}) \Big) \\ &+ \sum_{n=1}^{N} \sum_{k=1}^{K} (1 - \pi_{nk}) \log \Big(1 - \prod_{d=1}^{D} g(x_{nd} - t_{kd}^-) g(t_{kd}^+ - x_{nd}) \Big) \\ &+ \sum_{n=1}^{N} \sum_{k=1}^{K} \pi_{nk} \Big(-\frac{1}{2} \log |\boldsymbol{\Sigma}_k| - \frac{1}{2} (\mathbf{y}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{y}_n - \boldsymbol{\mu}_k) \Big).\end{aligned} \tag{5}$$

In the computation, we make use of the fact that $\mathbb{E}[z_{nk}] = \pi_{nk}$.

Maximization Step

After computing the expected value $Q(\mathbf{T}, \boldsymbol{\Theta})$ in Equation (5), we find the optimal $\mathbf{T}$ and $\boldsymbol{\Theta}$ the maximizes $Q(\mathbf{T}, \boldsymbol{\Theta})$ in the M step, such that

$$\widehat{\mathbf{T}}, \widehat{\boldsymbol{\Theta}} = \arg\max_{\mathbf{T},\boldsymbol{\Theta}} Q(\mathbf{T}, \boldsymbol{\Theta}) \tag{6}$$

By observing Equation (5), we conclude that we can maximize $\{\mathbf{t}_{kd}\}_{d=1}^{D}$ and $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$ independently for each $k \in \{1, \dots, K\}$.

The optimal $\{\mathbf{t}_{kd}\}_{d=1}^{D}$ is given by

$$\begin{aligned}&\Big\{ \widehat{\mathbf{t}}_{kd} \Big\}_{d=1}^{D} \\ &= \arg\max_{\{\mathbf{t}_{kd}\}_{d=1}^{D}} -\frac{1}{2}\alpha_t \sum_{d=1}^{D} (t_{kd}^- - \mu_{t_{kd}^-})^2 - \frac{1}{2}\alpha_t \sum_{d=1}^{D} (t_{kd}^+ - \mu_{t_{kd}^+})^2 - \frac{1}{2}\beta_t \sum_{d=1}^{D} (t_{kd}^+ - t_{kd}^-)^2 \\ &\quad + \sum_{n=1}^{N} \pi_{nk} \sum_{d=1}^{D} \Big(\log g(x_{nd} - t_{kd}^-) + \log g(t_{kd}^+ - x_{nd}) \Big) \\ &\quad + \sum_{n=1}^{N} (1 - \pi_{nk}) \log \Big(1 - \prod_{d=1}^{D} g(x_{nd} - t_{kd}^-) g(t_{kd}^+ - x_{nd}) \Big)\end{aligned} \tag{7}$$

Algorithm 1 The EM Algorithm

repeat

E step:

for $n \leftarrow 1$ **to** N **do**

Update the posterior distributions of $\mathbf{z}_n$ according to Equation (3) and (4).

end for

M step:

for $k \leftarrow 1$ **to** K **do**

Update $\{\mathbf{t}_{kd}\}_{d=1}^{D}$ according to Equation (7) using the BFGS algorithm.

Update $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ according to Equation (10).

end for

until Convergence

We solve this maximization problem with the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm [2]. With this algorithm, we need to make use of the gradient of Equation (7). Therefore, we compute the partial derivatives of Equation (7) with respect to t_{kd}^{+} and t_{kd}^{-}, respectively, as follows:

$$\begin{aligned}
&\frac{\partial Q(\mathbf{T},\boldsymbol{\Theta})}{\partial t_{kd}^{+}} \\
=& -\alpha_t t_{kd}^{+} + \alpha_t \mu_{t_{kd}^{+}} - \beta_t t_{kd}^{+} + \beta_t t_{kd}^{-} + \sum_{n=1}^{N} \frac{a\pi_{nk}}{1+\exp(a(t_{kd}^{+}-x_{nd}))} \\
&- \sum_{n=1}^{N}(1-\pi_{nk}) \frac{a\exp(-a(t_{kd}^{+}-x_{nd}))}{\exp(-a(t_{kd}^{+}-x_{nd}))+1} \frac{\prod_{d=1}^{D} g(x_{nd}-t_{kd}^{-})g(t_{kd}^{+}-x_{nd})}{1-\prod_{d=1}^{D} g(x_{nd}-t_{kd}^{-})g(t_{kd}^{+}-x_{nd})}
\end{aligned} \tag{8}$$

$$\begin{aligned}
&\frac{\partial Q(\mathbf{T},\boldsymbol{\Theta})}{\partial t_{kd}^{-}} \\
=& -\alpha_t t_{kd}^{-} + \alpha_t \mu_{t_{kd}^{-}} - \beta_t t_{kd}^{-} + \beta_t t_{kd}^{+} - \sum_{n=1}^{N} \frac{a\pi_{nk}}{1+\exp(a(x_{nd}-t_{kd}^{-}))} \\
&+ \sum_{n=1}^{N}(1-\pi_{nk}) \frac{a\exp(-a(x_{nd}-t_{kd}^{-}))}{\exp(-a(x_{nd}-t_{kd}^{-}))+1} \frac{\prod_{d=1}^{D} g(x_{nd}-t_{kd}^{-})g(t_{kd}^{+}-x_{nd})}{1-\prod_{d=1}^{D} g(x_{nd}-t_{kd}^{-})g(t_{kd}^{+}-x_{nd})}
\end{aligned} \tag{9}$$

We compute $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$ in closed form as follows:

$$\begin{aligned}
\boldsymbol{\mu}_k &= \frac{\sum_{n=1}^{N} \pi_{nk}\mathbf{y}_n}{\sum_{n=1}^{N} \pi_{nk}} \\
\boldsymbol{\Sigma}_k &= \frac{\sum_{n=1}^{N} \pi_{nk}(\mathbf{y}_n-\boldsymbol{\mu}_k)(\mathbf{y}_n-\boldsymbol{\mu}_k)^T}{\sum_{n=1}^{N} \pi_{nk}}
\end{aligned} \tag{10}$$

We repeat the E step and M step until the objective function defined in Equation (1) converges. The EM algorithm is summarized in Algorithm 1.

References

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[2] J. Nocedal and S. J. Wright. *Numerical Optimization*, chapter 6 Quasi-Newton Methods, pages 135 – 162. Springer, New York, 2nd edition, 2006.